Indian Journal of Science and Technology, *Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874, November 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Analyzing Diabetic Data using Classification Algorithms in Data Mining

K. Saravananathan[1] and T. Velmurugan[2]

[1]SRM Arts and Science College, SRM Nagar, Kattankulathur - 603203, Tamil Nadu, India; greatsaro@yahoo.co.in
[2]PG and Research Department of Computer Science, D.G. Vaishnav College, 833, E.V.R. Periyar High Road, Arumbakkam, Chennai - 600106, Tamil Nadu, India; velmurugan_dgvc@yahoo.co.in

Abstract

Backgrounds/Objectives: Huge medical datasets available in various data repositories which are used for real world applications. To visualize the useful information stored in data warehouses, the Data Mining (DM) methods are enormously utilized. One of such domain is medical domain, in which the function of DM approach raises speedy recovery of sickness over indications. On the way to categorize and predict symptoms in medicinal data, a variety of DM methods are utilized by different researchers. From many techniques of DM, classification is one of the main techniques. The classification techniques classify the unseen information in all areas including medical diagnostic field. The very dangerous disease in medicinal field is diabetes disease which is affected for many peoples in popular countries like India. **Methods/Statistical Analysis:** The impact of categorization is very important in authentic earth applications in all fields. To categorize the rudiments allowing to the applications of the elements during the predefined set of modules are used by classification methods. Very popular classification algorithms J48, Support Vector Machines (SVM), Classification and Regression Tree CART and k-Nearest Neighbor (kNN) for diabetic data are used for this research work. **Findings:** To discover the presentation of these classification methods, diabetic data as an input. For the most part, this research work is supported out to associate the techniques in the calculation of the presentation accurateness in diabetic data. The above mentioned techniques are used for diabetic data to categorize its accuracy in terms of its performance. **Methods:** The conclusion of this research work is choosing the top algorithm for the input data for the best classifier. **Applications/Improvements:** Some of other algorithms are analyzed using the same data set for the similar type of results is discussed in future. Also, some of the clustering algorithms are applied using the same data set to find highly affected diabetic patients.

Index Terms: CART Algorithm, Classification, J48 Algorithm, kNN Algorithm, SVM Algorithm

1. Introduction

The target of the information extracting method is to extract data from a dataset and make over it into a clear construction for additional use. This is a diagnostic method planned to scrutinized the information in seek of reliable patterns or organized associations connecting variables, and then to confirm the findings by applying the detected patterns. The focal point of this document is to concern a variety of categorization methods such as J48, kNN, CART and SVM.

As the commonness of diabetes is on the rise, there is a proportionate rise in the complications that are associated with diabetes and the illness has been the most deadly disease in the United States with no imminent cure in sight[1]. The diabetic sickness has number of side effects like eye disease, kidney failure, and additional complications. However, early detection of the disease and proper care management can make a difference[2]. It is the reasons sugar to build up in blood leading to complications like heart disease, stroke, blindness, kidney failure, nerve damage, and death. Regular Symptoms of Diabetes are increased thirst, increased urination, Weight loss, unsettled stomach or vomiting - Blurred vision, Slow-healing infections and weakness in men[3]. There are a variety of research work is carried out by many researchers based on

*Author for correspondence

the observed medical diabetes data. Some of such works are discussed hereafter.

Arvind Sharma and P.C. Gupta described that data mining can add by means of necessary benefits to the blood stockpile division. J48 method and WEKA software has been utilized for the entire research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9%[4]. The research is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using WEKA tool[5].

Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath[6], used cascading k-Mean and kNN algorithm for cataloging of diabetic patients in their paper. They classified diabetic patients by proposing results using kNN and k-Mean. Accuracy achieved by the proposed system is 82%. Hardik Maniya, Mosin I. Hasan, Komal P. Patel[7], have done the relative study of Naive Bayes Classifier and kNN for Tuberculosis, and justify the effectiveness of results using kNN can be further improved by increasing the number of data sets and for Naïve Bayesian classifier by increasing attributes or by selecting weighted features. W. Yu, and W. Zhengguo[8], have gave the investigational result shows the classification using traditional kNN algorithm produce normal evaluation value, with fulfillment rate of 75%. Y. Angeline Christobel, P. Sivaprakasam[9], the concert of classification calculate dregarding sensitivity, specificity and accuracy has been increased significantly in the case of proposed CkNN method.

In a research work of[10], Estebanez, Alter and Valls used genetic programming for classification tasks. The error rate for SVM is 22%, Simple Logistics is 22.14% and Multilayer perceptron is 23.31%. In another work some of the classification algorithms[11] are compared by utilizing matrix and classification accuracy. The 10-fold cross validation method was used by three different types of breast cancer databases and calculated the accuracy.

Jianchao Han[12] used type 2 diabetes data for his effort and the decision tree using WEKA has been used to put up the prediction model. The main element for his research was predicting the disease is the models of Plasma Insulin. Asma A. Aljarullah[13] in her research work J48 decision tree classifier was used. Using Diabetic data set was used to implement Association rule. B.M. Patil[14] finds out different range of accuracies using some of classification techniques on the diabetes dataset. Weighted least squares support vector machine based on quantum particle swarm optimization algorithm is used to development in the prediction accuracy.

E.G.Yildirim[15] in his research work the type 2 diabetes data set is used to predictive data mining and applied in dosage planning. Adaptive Neuro Fuzzy Inference System and Rough Set theory methods are coated by him. The main objective of G. Parthiban et al.[16] in their research work the prediction and changes of diabetic patient getting heart related problem. In their research they used Naive Bayes classifier method which gives the best possible prediction model.

The organization of this paper is formed as follows. Section II states some basic concepts of classification algorithms and its applications. Section III mentions the experimental results of the classification algorithms J48, CART, SVMs, and kNN for diabetes data set. Finally, the conclusion of this research work is given in section IV.

2. Materials and Methods

Data mining is the method of identifying, exploring and modeling huge amounts of data that discover unidentified Patterns or relationships that produce a correct result. There are various data classification algorithms available in DM. In which, some of the algorithms used for this research is discussed hereafter.

2.1. J48 Algorithm

The each and every phase of the information is to divide keen on slight subclasses to found on a decision. J48 inspect the standardized data grow that essentially the outcomes the dividing the data by selecting an element. To craft the conclusion, the element extreme regular data grow is utilized. Intense techniques bring to a halt if a subset related to the similar category in all the instances. J48 creates a result node use the projected values of the class. J48 be able to select particular attributes, lost attribute values of the information and contrary element values.

- First the leaf node is considered with the same set if the instances fit to the equal set.
- All attributes, the possible in sequence will be considered and the growth in information will be selected from the check on the attribute.
- The finest element will be identified derived from the current identification constraint.

2.2 Classification And Regression Tree (CART)

Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone together established an algorithm named as Classification and Regression Tree (CART) and formed a regular system for developing arithmetical models as of easy aspect data. CART is influential while it deals with information that is not fully completed, data with founded and contribution characteristics. The technique will study the some of examples in relation to the data description will tip to the data minimization and continues till some stop criteria is reached. At this point twofold splitting of attributes takes place. It affords an order of univariate binary decision.

Step1: To identify the method of splitting attribute is selected.

Step2: To determining upon what are the stop rules require to be in position.

Step3: How the nodes are used to divisions.

2.3 Support Vector Machines

Support Vector Machine (SVM) is founded on the idea of judgment planes that describe result limitations. The judgment idea is one that separates between a set of objects having a verity of class memberships. The regular SVM uses a collection of input data and predicts which of two possible classes comprises the input. An SVM representation of the examples as points in space, mapped so that the examples of the separated categories are divided by a clear gap that is as wide as possible.

2.4 k-Nearest Neighbor Algorithm

k-nearest neighbor algorithm is a simple technique that stores all available cases and classifies new cases based on a similarity measure. It is a type of lethargic knowledge where the function is only approximated nearby and the entire working out is deferred until classification. An entity is classified by the best part of its neighbors. k is always a positive integer. The correct classification is known because the neighbors are selected from a set of objects.

Step1: Determine k means the quantity of nearest neighbors.

Step2: Compute the distance between the query instance and all the training samples.

Step3: The distance of all the training samples are sorted and nearest neighbor based on the k minimum distance is determined.

Step4: Get all the categories of the training data for the sorted value which falls under k.

Step5: Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

2.5 Statistical Measures

For the calculation of the section of the predicted positive cases the below mentioned formulas are used. Precision P using TP is True Positive Rate and FP is False Positive Rate and they defined as,

$$\text{Precision P} = \frac{TP}{TP + FP} \tag{1}$$

The proportion of positive cases that were correctly identified are known as True Positive Rate (TPR). It is calculated as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

Where FN = False Negative Rate

In this research work, there are three measures used. Correctly classified instances are properly classified by any classification technique. Accuracy is calculated by an exact value.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

The mentioned rule for the accuracy calculation the above mentioned formula is used with TN = True Negative.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \tag{4}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \tag{5}$$

The F-Measure can be computed as some average of the information retrieval precision and recall metrics.

$$\text{F} = \frac{2 * Recall * Precision}{Precision + Recall} \tag{6}$$

Kappa Statistics evaluate amount of concurrence between two sets of classified data. Kappa result varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement.

$$k = \frac{p(a) - p(e)}{1 - p(e)} \tag{7}$$

Where p(a) = percentage of agreement, p(e) = chance of agreement. The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|fi - yi| = \frac{1}{n}\sum_{i=1}^{n}|ei| \quad (8)$$

The root mean squared error RMSE ei of an individual program i is evaluated by the equation:

$$ei = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(P(i,j) - Tj\right)^2} \quad (9)$$

Where, P(i, j) = the value predicted by the single program, i = fitness case, Tj = the target value for fitness case j. Relative Absolute Error Ei is calculated by the expression:

$$Ei = \sum_{j=1}^{n}|Pij - Tj| \Big/ \sum_{j=1}^{n}|Tj - T''j| \quad (10)$$

Where, Pijis the value predicted by the single program i for sample case j; Tjis the final value for sample case j; and Tis known by the formula:

$$T'' = \frac{1}{n}\sum_{j=1}^{n} Tj \quad (11)$$

The relative squared error (RSE) can be compared between models whose errors are measured in the different units.

3. Experimental Results

The diabetes data set is used for this research work. This data set has 10 attributes namely age, plasma glucose fasting, plasma glucose post, urea, Creatinine, Sodium, Potassium, HBA1C, Name and Sex for 545 patients. The coreplan of this research work is to evaluate the performance of classification methods for diabetes data based on the numerical input constraints. The data are evaluated using J48, CART, SVM, and kNN algorithms. For the classification all the values of ten attributes chosen and accepted for pre-processing. A relative analysis of classification accurateness using J48, CART, SVM, and kNN technique is accepted in this work.

Totally, 545 patients' data is collected from a private medical Diabetic center. In which, there are 366 male and 179 female patients whose age between 40 and 60 years. This research work mainly discusses about the accuracy of classification algorithms compared with the execution time and error rate using WEKA software. The various attributes in the diabetes data set are described in table 1. In the given data, the attribute sex has two classes (Male/ Female), Plasma glucose is in two ranges Fasting and Post prandial and other values as per the blood samples. The normal accepted ranges of attribute values for the attributes numbered from 3 to 9 are given in the end of the table.

Figure 1, shows the circulation of data set values based on the values of Plasma Glucose (Fasting) ranges from diabetes dataset. Also, the circulation of data set values Plasma Glucose (Post Prandial) ranges is shown in Figure 2. The investigational results of basic classifiers are described in this segment. Based on the values of Plasma Glucose (Fasting) and Plasma Glucose (Post Prandial) are taken for the analysis in this research work for classification. The minimum value stored in Fasting glucose is 76 and the maximum value is 184. Similarly, the minimum and maximum values available for post prandial glucose are 106 and 202 respectively. This information is depicted in the respective figures 1 and 2. To categorize the diabetes data suitably since the working out data set, the error rates and accuracy are calculated using classifiers. The accuracy of J48 method found to be 67.16 %, CART is 62.29%, Support Vector Machines is 65.05% and kNN 53.39%. The results of various measures are given Table 2.

By the use of addition of true positive and true negative continued by the division of all possibility values the

Table 1. Description of the Data Set

S. No.	Variables	Reference Value
1	Sex	Male / Female
2	Age	Between 40 and 60
3	Plasma Glucose Fasting (PGF)	75 – 115
4	Plasma Glucose Post Prandial (PGP)	75 – 140
5	Serum Urea	10 – 50
6	Serum Creatinine	0.6 – 1.1
7	Serum Sodium	130 – 145
8	Serum Potassium	3.5 – 5.0
9	HBA1C	4.0 – 6.0

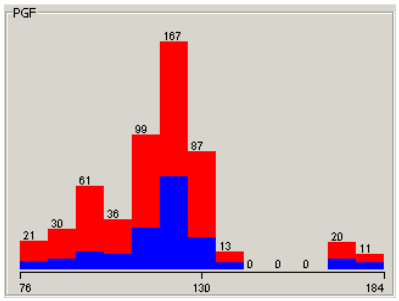

Figure 1. Data distribution for Plasma Glucose (Fasting).

accuracy is calculated. Accuracy is measured and created using 10fold cross validation method. Tenfold cross-validation is the general way of calculating the error rate of a learning scheme on a particular dataset. In 10-fold cross validation method, there are ten equal sized partitions of the data set.

By using this method, huge test data sets produces a good assessment of the classifier's performance and small training data sets result is in a poor classifier. The Table 3 gives the error values of the taken four classification techniques. Table 4 point out the accuracy results of the rightly classified and wrongly classified instances of the classification algorithm for the diabetes dataset. In Figure 3 shows the comparison of performance accuracy of the various techniques mentioned, it contains correctly classified and incorrectly classified instances. The comparison performance accuracy of the different methods represented as shown in figure 4, which contains only correctly classified instances. Main intention of this work is to match up to different classification algorithms accuracies.

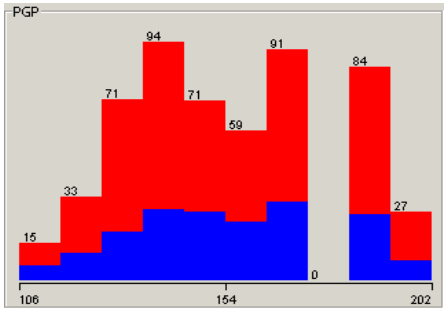

Figure 2. Data distribution for Plasma Glucose (Post Prandial).

Table 2. Results of Various Measures

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
J48	0	0	0	0	0	0.494	Female
	1	1	0.672	1	0.804	0.494	Male
	0.672	0.672	0.451	0.672	0.54	0.494	Weighted Average
CART	0	0	0	0	0	0.49	Female
	1	1	0.623	1	0.768	0.49	Male
	0.623	0.623	0.388	0.623	0.478	0.49	Weighted Average
SVMs	0.006	0.012	0.2	0.006	0.011	0.497	Female
	0.988	0.994	0.655	0.988	0.788	0.497	Male
	0.65	0.657	0.499	0.65	0.521	0.497	Weighted Average
kNN	0.296	0.35	0.293	0.296	0.294	0.485	Female
	0.65	0.704	0.654	0.65	0.652	0.485	Male
	0.534	0.588	0.535	0.534	0.535	0.485	Weighted Average

Table 3. Error Reports

STATISTIC	J48	CART	SVMs	kNN
Kappa statistic	0	0	-0.008	-0.0535
Mean absolute error	0.4411	0.4698	0.3495	0.4625
Root mean squared error	0.4697	0.4847	0.5912	0.6759
Relative absolute error	99.9458 %	99.9595 %	77.4368 %	104.7935 %
Root relative squared error	99.9999 %	99.9999 %	124.4752 %	143.9201 %

Table 4. Performance Accuracy

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
J48	67.156	32.844
CART	62.2857	37.7143
SVM	65.0485	34.9515
kNN	53.3945	46.6055

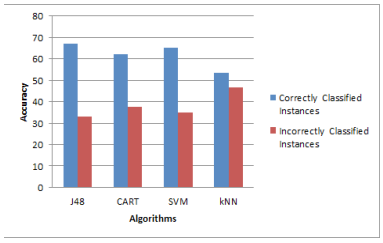

Figure 3. Performance comparison of algorithms.

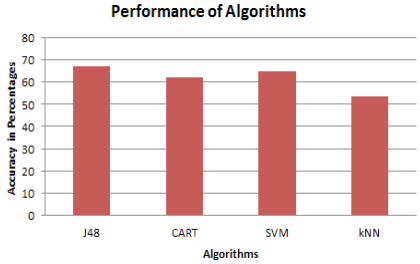

Figure 4. Accuracy of Algorithms.

4. Conclusion

In this research work, the frequently used classification techniques J48, CART, SVMs, and kNN are analyzed, on the medical dataset to find the optimal solution for Diabetes. The performance indicators accuracy, specificity, sensitivity, precision, error rate are calculated for the given dataset. Accusation beside with a proper data preprocessing technique can get better the accuracy of the classifier. The function of data normalization had noticeable impact on categorization performance and considerably enhanced the performance of J48. The performance of kNN algorithm has minimum accuracy. Based on the parameters taken for analysis, the performances of the four algorithms are analyzed. The results show that the performance of J48 technique is significantly superior to the other three techniques for the classification of diabetes data. To improve the overall accuracy, it is necessary to use more data set with large number of attributes and use the best feature selection method in future. Future works may also include hybrid classification models by combining some of the data mining techniques.

5. References

1. Iyer Aiswarya, Jeyalatha S and Sumbaly Ronak. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process. 2015; 5:1-14.
2. Velide Phani Kumar and Velide Lakshmi. A Data Mining Approach for Prediction and Treatment of diabetes Disease. International Journal of Science Inventions Today. 2014; 3:73-9.
3. Sanakal Ravi and Jayakumari T. Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine. International Journal of Computer Trends and Technology. 2014; 11:94-8.
4. Sharma Arvind and Gupta PC. Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool. International Journal of Communication and Computer Technologies. 2012; 01:6-10.
5. Yasodha P, Kannan M. Analysis of a Population of Diabetic Patients Databases in WEKA Tool. International Journal of Scientific & Engineering Research. 2011; 2:1-5.
6. Karegowda Asha Gowda, Jayaram MA, Manjunath AS. Cascading K-means Clustering and k Nearest Neighbor Classifier for Categorization of Diabetic Patients. International Journal of Engineering Advanced Technology. 2012; 1:147-51.
7. Maniya Hardik, Mosin I Hasan, Komal P Patel. Comparative study of Naive Bayes Classifier and kNN for Tuberculosis. International Journal of Computer Applications. 2011; p. 22-6.
8. Yu W and Zhengguo W. A Fast kNN algorithm for text categorization. Hong Kong: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. 2007; 50:3436-41.
9. Angeline Christobel Y, Sivaprakasam P. A New Class wise k Nearest Neighbor (CkNN) Method for the Classification of Diabetes Dataset. 2013; 2:396-400.
10. Estebanez C, Aler R and Valls M. Genetic Programming Base Data Projections for Classification Tasks. World Academy of Science, Engineering and Technology. 2005; p. 56-61.
11. Salama GI, Abdelhalim MB, Zeid MA. Experimental comparison of classifiers for breast cancer diagnosis. International Conference on Computer Engineering & Systems. 2012; 98:180-5.
12. Ianchao Han J, Juan C Rodriguze, Beheshti Mohsen. Diabetes Data Analysis and Prediction model discovery. Second International conference on future generation communication and networking. 2011; p. 96-9.
13. Asma A Aljarullah. Decision tree discovery for the diagnosis type-2 diabetes. International conference on innovation in information technology. 2011; p. 303-7.
14. Patil BM, Joshi RC, Toshniwal Durga. Hybrid prediction model for type-II diabetic patients. Expert Systems with Applications. 2012; p. 8102-8.
15. Yildirim EG, Karachoca A and Uear T. Dosage Planning for diabetes patients using data mining methods. Procedia Computer Science. 2011; p. 1374-80.
16. Parthiban G, Rajesh A, Srivatsa SK. Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method. International Journal of Computer Applications. 2011; 24:7-11.